SEGUNDA PROVA DE ESTATÍSTICA COMPUTACIONAL - 14/06/2021

A ave Petorus Franklinulata pode ser encontrada em alguns países nórdicos. O conjunto de dados `aves.txt` apresenta informações sobre a envergadura da asa (em cm), o peso (em gramas) e a localidade (país onde vive) de uma amostra de 150 aves dessa espécie. Todas as questões abaixo referem-se a esse conjunto de dados.

Questão 1. (a) Leia o arquivo `aves.txt`. Em seguida, comece a analisar os dados a partir das funções `head, tail, str, summary`.

(b) Conforme pôde ser observado na parte (a), os dados estão ordenados a partir da variável `local`. Embaralhe o conjunto de dados de modo a deixar a ordenação aleatória.

(c) Determine a média e o desvio padrão das variáveis `comprimento_asa` e `peso` para as espécies de cada localidade (variável `local`). Comente os resultados encontrados.

(d) Plote o gráfico de `comprimento_asa` versus `peso`. Nesse gráfico, cada localidade deve ser representada por uma cor diferente.

(e) Divida o conjunto de dados em dois: um para treino e outro para teste. O conjunto de treinamento deve conter 80% dos dados do conjunto inicial.

(f) A partir de uma inspeção gráfica (parte (d)), crie uma árvore de decisão para classificar uma ave de acordo com sua localidade, isto é, as variáveis de entrada da árvore devem ser `comprimento_asa` e `peso` e a variável de saída (classificação) deve ser `local`.

(g) Calcule a sua taxa de acerto para o modelo construído em (f) utilizando nesse cálculo o conjunto de teste. Em seguida, construa a matriz de confusão e, por fim, comente os resultados encontrados.

Questão 2. (a) Divida o conjunto de dados `aves.txt` em três data frames: um para cada localidade. Em seguida, calcule o coeficiente de correlação entre as variáveis `comprimento_asa` e `peso` em cada um dos data frames. Em qual localidade as variáveis estão mais correlacionadas linearmente?

(b) Sejam X e Y dois vetores de tamanho n. A equação da reta de regressão para a variável independente X e para a variável dependente Y é:

$$\hat{Y} = mX + b$$

Construa uma função cuja entrada sejam vetores `X` e `Y` (ambos de tamanho n) e cuja saída seja um vetor de tamanho 2 em que a primeira posição é m e a segunda posição é b.

(c) Utilize o conjunto de dados da localidade em que as variáveis `comprimento_asa` e `peso` estão mais correlacionadas para determinar a reta de regressão linear simples entre essas duas variáveis. Considere que `comprimento_asa` é a variável independente. Para a obtenção da reta, utilize a função construída em (b).

(d) Uma variação de 0.5 cm no comprimento da asa provocaria uma variação de quantos gramas no peso da ave? Por que?

(e) O Excelentíssimo Senhor David Attenborough acredita que podemos utilizar esse modelo para prever o peso de uma ave que tem uma envergadura de asa de 23 cm. O Excelentíssimo Senhor David Attenborough está certo? Justifique sua resposta. Se sua resposta for afirmativa, qual seria, de acordo com o modelo, o peso dessa ave?

Para a Questão 2b, os valores de m e b são dados por:

Teorema 1. *A equação de uma reta de regressão para uma variável independente X e uma variável dependente Y é:*

$$\hat{Y} = mX + b$$

em que $\hat{Y}$ é o valor previsto de Y para um dado valor de X. A inclinação m e o intercepto em Y, b, são dados por:

$$\begin{aligned} m &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \\ b &= \overline{y} - m\overline{x}, \end{aligned}$$

em que $\overline{y}$ é a média dos valores de Y no conjunto de dados, $\overline{x}$ é a média dos valores de X e n é o número de pares de dados.